Indexer, comparer, apparier des textes et leurs résumés : une exploration.

Martine Cadot, Sylvain Aubin, Alain Lelu

Indexer, comparer, apparier des textes et leurs résumés : une exploration.

Martine Cadot(1), Sylvain Aubin (2), Alain Lelu (1, 3, 4)

(1) LORIA, Campus scientifique, BP 239, 54506 Vandoeuvre cedex – France
(2) Diatopie SA, 27 Bd. St. Martin 75003 Paris
(3) LASELDI, 30 rue Mégevand – 25030 Besançon cedex
(4) ISCC, 20 r. Berbier-du-Metz 75013 Paris
martine.cadot@loria.fr, sylvain.aubin@diatopie.com
alain.lelu@univ-fcomte.fr

Résumé : Nous présentons ici la démarche qui nous a valu un score de 100% de réussite au défi DEFT 2011 dans la tâche d'appariement de résumés avec des articles dépourvus d'introduction et de conclusion : nous avons testé plusieurs types d'indexation et de distance résumé-texte, et mis au point une méthode d'appariement, en univers fermé, robuste et sans nécessité d'information extérieure. En combinant quatre variantes de la distance de compression, indépendante de la langue et du type de codage, elle permet d'atteindre 93% ; les 100% sont atteints avec la distance de Hellinger appliquée à des textes indexés par des noms lemmatisés et des termes composés, distance qui surpasse ici la classique TF-IDF. Nous suggérons son application en univers ouvert, avec plus de textes que de résumés, et des résumés sans texte.

Abstract: We develop here the approach which enabled our 100% success score in the DEFT 2011 challenge for the task of mating abstracts to their respective papers in the domain of social sciences and humanities. These papers were preliminarily reduced by their introduction and conclusion: we have tested several indexing methods and inter-text distance formulas between abstracts and papers. We then settled a mating method specific for the case when their relation is bijective. This method proved robust, and needs no external information source. Using the normalized compression method resulted in a 93% success score; using a crude lemmatizer/tagger, a key-phrase extractor and the Hellinger distance resulted in a 100% score. This distance proved to behave slightly better than the Salton's TF-IDF, and to be more fitted to incremental "open" corpuses processing, a more realistic mating problem for which we suggest a variant of our method.

Mots-clés : similarité textuelle, distance de compression, distance de Hellinger, TF-IDF, lemmatisation, extraction de termes composés, indexation, étiquetage morpho-syntaxique.
Keywords: text similarity, compression distance, Hellinger distance, TF-IDF, lemmatization, key-phrase mining, indexing, morpho-syntactic tagging.

1 Introduction : le problème à résoudre

Le défi DEFT 2011 comportait deux tâches à résoudre, un ensemble de textes anciens à dater d'une part, un ensemble de résumés d'articles de sciences humaines à apparier à leurs textes d'origine d'autre part. Nous avons choisi de nous concentrer sur cette dernière tâche, qui comportait deux « pistes », la plus difficile consistant à apparier les résumés avec des textes amputés de leur introduction et de leur conclusion, alors que l'autre dévoilait les textes complets. Ce défi a présenté pour nous l'intérêt de comparer quantitativement diverses solutions possibles à chaque étape de la chaîne des traitements, à savoir plusieurs méthodes d'indexation, diverses distances entre textes, et de mettre au point une procédure d'appariement, pour finalement évaluer les meilleures combinaisons de ces éléments. En effet, à l'instar de l'option choisie pour un autre défi DEFT (Lelu et al., 2006), nous avons pris le parti de multiplier les points de vues sur les données, ainsi que leurs critères de comparaison, en créant plusieurs chaînes de traitement les plus « orthogonales » possibles. De deux choses l'une : ou bien aucune de ces chaînes ne donnait zéro erreur sur l'ensemble d'apprentissage, et leur combinaison avait toute chance d'améliorer chacun de leurs résultats individuels, par exemple par une procédure de vote pondéré ; ou bien l'une d'elles ne donnait aucune erreur, et nous n'avions aucun élément pour douter de ses seuls résultats dans la phase de test. On verra que c'est cette dernière éventualité qui s'est produite.

Le défi définissait un univers de textes « fermé », au sens où à chaque résumé correspondait un texte, et vice-versa. Comme beaucoup de problèmes réels d'appariement de textes se situent en univers « ouvert », où il s'agit de trouver un ou des texte(s) les plus proche(s) d'un texte donné au sein d'une large collection, nous avons évoqué en conclusion les modifications à apporter à notre méthode d'appariement pour laisser de côté l'information « correspondance biunivoque entre résumés et textes ».

2 Les données

Les données proposées sont tirées de revues de sciences humaines en ligne dans la plateforme québécoise Erudit (*http://www.erudit.org/*). L'ensemble d'apprentissage rassemblait 300 textes des 5 revues *Anthropologie et Sociétés, Études internationales, Études littéraires, Meta, Revue des sciences de l'éducation* et les résumés correspondants. L'ensemble de test comportait en plus des textes et résumés de la revue *Philosophiques*. La difficulté de la tâche tenait en premier lieu pour la piste 1 à l'absence d'introduction et de conclusion, dont les éléments de synthèse sont souvent repris pour tout ou partie dans les résumés. Mais elle tenait surtout à la présence de certains résumés, parfois très succincts, pour de longs textes écrits par le même auteur, sur le même sujet, dans la même revue ! Seuls des spécialistes en traductologie, ou en histoire de la philosophie, par exemple, pouvaient être capables de les apparier correctement au fil d'une lecture attentive.

Il y avait quelques difficultés ponctuelles en plus : un résumé en grec dans l'ensemble d'apprentissage – nous nous sommes contentés d'une « Google-traduction » pour l'intégrer - , un texte vide dans l'ensemble de test – la difficulté était facilement résolue compte tenu de la bi-univocité des appariements.

3 Indexations et distances entre textes

Nous avons opté pour trois définitions contrastées de distances entre textes : deux font appel à un pré-traitement d'indexation, soit basique, soit élaboré, l'autre se veut la plus brute possible et indépendante de la langue. Pour cette dernière l'utilisation de N-grammes de caractères était un bon candidat, d'autant plus que nous l'avions déjà employée avec succès (Delprat et al., 2011). Mais nous avons voulu tester les possibilités d'une autre méthode, encore plus « aveugle », qui ne demande pas même le minimum de pré-traitements nécessaires pour les N-grammes : la distance de compression (Cilibrasi, 2003), indépendante de la langue et du codage textuel utilisé, qui ne nécessite pas même d'ouvrir chaque fichier texte. Nous décrivons tout d'abord cette dernière.

3.1 Distance de compression

La distance de compression normalisée entre deux chaînes de caractères x et y est définie comme suit par les auteurs cités :

$D_c(x,y) = [Z(xy) – \min(Z(x), Z(y))] / \max(Z(x), Z(y))$

où Z(x) est la longueur du texte x après compression sans perte, et xy désigne la concaténation des textes x et y. Nous avons utilisé ici le compresseur zip.exe du groupe Info-Zip (version 2.32, en ligne de commande) implantant l'algorithme de déflation (cf. *http://www.info-zip.org/*). A noter que du fait du caractère séquentiel et non-optimal de cet algorithme de compression sans perte, D(x,y) diffère en général de D(y,x) quand les longueurs des deux textes diffèrent fortement, ce qui est le cas ici. Nous avons utilisé les deux configurations « résumé puis texte » et « texte puis résumé » sur les ensembles d'apprentissage et de test.

Compte tenu du déséquilibre entre les tailles des résumés (notés r) et des textes t la distance ci-dessus revient à : $D_{c1}(r,t) = [Z(rt) –Z(r)] / Z(t)$. Dans le cadre de notre volonté de multiplier les points de vues, ne serait-ce que dans le cas d'une seule méthode, nous avons également utilisé la dissimilarité

$D_{c2}(r,t) = [Z(rt) –Z(t)] / Z(r)$ normalisée par rapport au seul résumé, a priori plus « sensible » et sujette à de fortes variations entre 0 et 1.

L'avantage de ce type de distance est qu'il ne nécessite absolument aucun pré-traitement et est indépendant de la langue et du type de codage des caractères. L'inconvénient de cet avantage est qu'il est une boîte noire qui ne montre pas le pourquoi de ses résultats, contrairement aux distances vectorielles qui peuvent, si on le désire, expliciter les descripteurs communs à deux textes. Son inconvénient pratique est qu'il nécessite |R|*|T| opérations de concaténation de fichiers, où |R| et |T| sont respectivement les nombres de résumés et de textes, soit ici 90 000 opérations pour l'ensemble d'apprentissage.

3.2 Indexation basique par formes brutes

Nous avons choisi, comme indexation de référence à titre d'intermédiaire entre l'absence d'indexation décrite ci-dessus et l'indexation évoluée de type morphosyntaxique, les chaînes de caractères séparées par des espaces (ordinaires ou insécables) ou un séparateur parmi les caractères de la liste : *.,;:?!'-«»()[]{}*. Nous appellerons « formes brutes » ces chaînes de caractères.

L'avantage de cette approche est d'être la plus simple pour obtenir des vecteurs-documents. Elle n'est opératoire que pour les langues à caractères alphabétiques et non agglutinantes, dans lesquelles il est trivial de séparer les formes. Pour les langues agglutinantes ou sans séparateurs triviaux de mots, la technique des N-grammes de caractères s'impose, ce qui n'est pas le cas ici. Nous avons obtenu un dictionnaire de 26 266 formes brutes, hors hapax, pour l'ensemble d'apprentissage des textes tronqués, de 21 270 pour l'ensemble de test correspondant.

3.3 Indexation élaborée par lemmes et expressions composées

Nous avons utilisé à cet effet le logiciel NeuroNav (commercialisé par Diatopie SA, et décrit dans *http://webu2.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm*) qui comporte, outre une interface de consultation d'une base textuelle et navigation dans les « axes obliques » synthétiques qui en sont extraits, deux modules utilisés pour la présente étude :

- Un module d'étiquetage grammatical et lemmatisation, pour le français et l'anglais, qui étiquette de façon minimale en trois catégories : les adjectifs, les verbes et les substantifs. Une quatrième catégorie, celle des particules syntaxiques, est éliminée dans la phase d'indexation. Chaque forme reconnue dans un dictionnaire de formes donne lieu à un lemme étiqueté, selon une heuristique basique, mais globalement efficace : le lemme et l'étiquette sont ceux dont l'emploi est le plus fréquent dans la langue générale. Les formes absentes du dictionnaire sont étiquetées substantif, par défaut.

- Un module d'extraction d'expressions composées (jusqu'à 6 composants), à partir de patrons syntaxiques, comme *Subst* de *Subst*, *Subst Adj*, *Subst* en *Subst*, etc. Des listes d'exceptions permettent de filtrer des expressions de la rhétorique courante, comme *nombreuses façons* en français. Pour l'anglais NeuroNav est orienté davantage vers l'analyse de bases documentaires ou journalistiques que de textes littéraires, et des expressions telles que *excellent results* ou *given context* sont éliminées. Le but est de privilégier la précision sur le rappel, pour ne pas troubler l'utilisateur par la présence de nombreux termes d'indexation non pertinents. Une étude sur une

petite base de résumés dans le domaine génomique indexés à la main a montré une précision supérieure à 95%, pour un rappel de termes composés corrects de l'ordre de 50%. Ce taux de rappel médiocre, dû aussi pour beaucoup au vocabulaire très particulier du domaine, est rendu acceptable par la redondance présente dans tout texte, et dans notre cas aucun des 300 résumés, puis des 200 de la phase de test, aussi court soit-il, ne s'est vu attribuer aucun terme, malgré l'élimination des mots et expressions hapax, inutiles pour les comparaisons résumés/textes à effectuer.

Au final, l'indexation des résumés et textes tronqués de l'ensemble d'apprentissage s'est traduite par un total de 23 331 termes lemmatisés différents, hors hapax, dont 3771 adjectifs, 2062 verbes, 10 412 substantifs et 7286 expressions composées. L'ensemble de test a fourni de son côté 16 365 termes.

3.4 Distances entre textes indexés

Un ensemble de textes indexés peut être représenté par un ensemble de vecteurs dont les composantes sont le nombre d'occurrences du descripteur i dans le texte t (modèle « sac de descripteurs »). L'indicateur de similarité le plus utilisé pour les textes (Banerjee et al., 2005) est le cosinus, obtenu par simple produit scalaire entre deux vecteurs-textes normalisés $<\mathbf{v}_t, \mathbf{v}_{t'}>$, qui permet de comparer des textes de longueurs différentes. La façon de normaliser, ainsi que la pondération éventuelles des composantes, définissent une large palette de possibilités, dont nous retiendrons deux parmi les plus utilisées, et une moins connue dont nous avons déjà montré les mérites (Lelu, 2003). Pour chacun de ces cosinus, nous définissons de façon homogène la distance entre deux documents comme l'arc de leur cosinus, ce qui évite certaines variantes qui peuvent avoir une influence sur les résultats en univers fermé : en effet certains auteurs définissent la « distance du cosinus » comme 1-cos, alors que plus rigoureusement la distance sur l'hypersphère est définie par la fonction ArcCosinus, et la distance de la corde, qui lui est équivalente pour toutes opérations de tri et seuillage, s'écrit $\sqrt{2} *(1-\cos)^{1/2}$ Nous définissons ci-après trois transformations du vecteur-document brut produisant des vecteurs normalisés. La distance correspondante entre deux textes sera définie dans tous les cas comme l'arc de leur cosinus, cosinus obtenu par produit scalaire entre ces deux vecteurs-textes normalisés.

- **Distance euclidienne entre vecteurs-textes normalisés**

Chaque composante k_{it} (occurrence du mot i dans le texte t) du vecteur-texte $\mathbf{v}_t$ est divisée par la norme euclidienne de ce vecteur :

$\{ k_{it} \} \rightarrow \{ k_{it} / \| \mathbf{v}_t \| \}$ où $\| \mathbf{v}_t \| = (\Sigma_i k_{it}^2)^{1/2}$

Dans ce cas le vecteur de longueur 1 obtenu est colinéaire au vecteur d'origine.

- **Distance « TF-IDF »**

Dans cette distance, très utilisée dans les applications de recherche de l'information, les occurrences du mot i (*Term Frequency*) sont pondérées de façon à diminuer l'importance des termes fréquents (*Inverse Document Frequency*) :

$\{ k_{it} \} \rightarrow \{ k_{it} \log(N/ n_{i.}) \}$ où $n_{i.}$ est le nombre de présences (et non d'occurrences) du mot i dans l'ensemble des unités textuelles, et N le nombre de textes. Les composantes de ce nouveau vecteur sont alors divisées par sa norme. Le vecteur de longueur 1 obtenu n'est plus colinéaire au vecteur d'origine.

A noter que cette distance est mal adaptée à l'arrivée incrémentale de nouveaux textes, sauf à supposer une certaine stabilité dans la distribution globale des présences de termes $n_{i.}/N$, ces derniers pouvant modifier le système des distances entre textes utilisé auparavant.

- **Distance de Hellinger**

Cette distance semble très proche de la distance euclidienne sur la sphère, voire redondante avec elle, puisque chaque vecteur-texte est normalisé sans intervention des fréquences ou présences globales de descripteurs $k_{i.}$ ou $n_{i.}$. Chaque vecteur-texte est normalisé comme suit :

$\{ k_{it} \} \rightarrow \{ (k_{it} / k_{.t})^{1/2} \}$

Elle est également adaptée au cas de l'analyse d'un corpus de façon incrémentale, au fil de l'arrivée de nouveaux documents indexés, qui ne changent pas les valeurs des composantes des vecteurs précédents. A noter que chaque vecteur de longueur 1 obtenu $\mathbf{v}_t$ n'est plus colinéaire à son vecteur d'origine.

La distance de Hellinger D_H est la longueur de la corde correspondant à l'angle $(\mathbf{v}_t, \mathbf{v}_{t'})$ - égale au plus à 2 quand ces deux vecteurs normalisés sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux.

$D_H(t,t') = \sqrt{2} *(1-<\mathbf{v}_t, \mathbf{v}_{t'}>)^{1/2}$

Plusieurs propriétés théoriques la rendent intéressante de notre point de vue :

- Elle est particulièrement adaptée aux « données directionnelles » (Banerjee et al., 2005) que sont les données textuelles, pour lesquelles seuls sont pertinents les angles entre vecteurs.

- Elle est liée à la mesure du gain d'information de Renyi d'ordre ½ (Renyi, 1966) apporté par une distribution $\mathbf{x}_q$ quand on connaît la distribution $\mathbf{x}_p$:

$$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = - 2 \log_2 (\cos(\mathbf{z}_p, \mathbf{z}_q)) = - 2 \log_2 (1 - D_H^2/2)$$

- et surtout (Escofier, 1978) et (Domengès et Volle, 1979) ont montré qu'elle satisfaisait à la même propriété d'équivalence distributionnelle que la distance du khi-deux utilisée en Analyse Factorielle des Correspondances : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, dans le cas où les descripteurs sont des mots et les unités décrites des textes, cette propriété assure la stabilité du système des distances entre textes au regard de l'éclatement ou du regroupement de mots de distributions proches. Ceci peut expliquer la considérable supériorité de ses performances qu'on constatera plus bas par rapport à la distance euclidienne sur la sphère, et qui confirme des constats faits par ailleurs – ex. : (Legendre et Gallagher, 2003).

4 Notre méthode d'appariement

4.1 La première étape de notre méthode d'appariement

Nous décidons d'apparier résumés et textes à l'aide d'une méthode s'inspirant des voisins réciproques dont le principe est : on affecte à un résumé *r1* le texte *t* le plus proche, puis on affecte au texte t le résumé *r2* le plus proche, et si le résumé r1 se trouve être le résumé r2, on estime avoir réussi l'appariement entre le résumé et le texte. Nous allons illustrer ce principe basique sur les distances obtenues par compression entre les résumés et les textes, en utilisant la variante : distance 2, configuration « résumé puis texte ». Pour cela nous avons créé une première liste de couples (résumé, texte) candidats à l'appariement en cherchant pour chaque résumé le texte le plus proche, et une deuxième liste de couples en cherchant pour chaque texte le résumé le plus proche, afin de les confronter pour la recherche des voisins réciproques. Mais nous constatons un grand nombre de « doublons ». En effet, parmi les 300 couples de la deuxième liste, le même résumé (r90) était affecté à 295 des 300 textes alors qu'il aurait dû n'être affecté qu'à un seul texte, le résumé r182 était affecté à 3 textes, et les 2 derniers textes de la liste produisaient les 2 couples (r103, t235) et (r145, t234), associations qui se sont avérées justes. La première liste était plus intéressante, car seulement 22 des 300 textes se trouvaient être les plus proches de plusieurs résumés (2 ou 3), ce qui laissait 252 associations possibles entre un résumé et un texte. Parmi celles-ci, une seule était fausse. En combinant les deux listes pour obtenir les voisins réciproques, nous n'avions plus qu'un couple. Devant ce résultat, nous avons décidé de garder les voisins réciproques, qui ont formé les appariements de qualité 1, mais aussi les couples de « qualité 2 », c'est-à-dire présents dans une des deux listes seulement *à condition qu'ils ne vienne pas en contradiction avec ceux de l'autre liste*, en définissant les couples en contradiction des couples qui ont un élément commun, comme (r, t') et (r, t''), ou (r', t) et (r'', t). Ainsi les deux listes qui avaient l'une deux couples et l'autre 252 couples ont été combinées en une liste de 253 couples : 1 couple de qualité 1 (voisins réciproques), et 252 de qualité 2. Sur les 300 couples attendus, on en a trouvé 252 justes et 1 faux, donc un taux de reconnaissance correcte de 84%. On peut voir dans le *Tableau 1* le résultat de cette première étape pour les 4 variantes de la méthode.

Variante	Distance	Ordre	Taux de reconnaissance (bien reconnus/tous)	Nombre d'erreurs
1	1	résumé, texte	84,00%	1
2	1	texte, résumé	55,33%	0
3	2	résumé, texte	41,67%	0
4	2	texte, résumé	37,00%	1

Tableau 1 : Résultats sur l'ensemble d'apprentissage

Les résultats pour les 3 autres variantes de la méthode de compression sont moins bons, mais les 4 variantes ont en commun leur très petit nombre d'erreurs (0 ou 1), ce qui va permettre d'augmenter leur performance par combinaison, comme indiqué dans le Tableau 2. On a réparti dans ce tableau les 300 associations attendues, correspondant donc aux 300 résumés, en indiquant en colonne le résultat final (-1 : appariement faux, 0 : non apparié, 1 : appariement juste) et en ligne la qualité (-1 : appariements proposés en contradiction, 1 : appariement avec accord des 4 variantes, 2 : accord de 3 variantes, 3 : accord de 2 variantes, 4 : 1 variante , et 5 : aucune variante donc pas d'appariement proposé).

Nombre d' associations	Validation			
qualité	-1	0	1	Total
-1		1*		1
1			46	46
2			68	68
3			100	100
4	1		64	65
5		20		20
Total	1	21	278	300

Tableau 2 : Combinaison des résultats des 4 variantes sur l'ensemble d'apprentissage

On a indiqué par une étoile dans le tableau le nombre d'appariements proposés par plusieurs variantes qui ont produit une contradiction. Il y en a un seul. Il est situé sur la ligne de qualité -1 et dans la colonne de validation 0 car il n'a pas abouti à une association. Il s'agit du résumé 67, qui a été associé au texte 38 dans les deux premières variantes, à aucun dans la troisième, et au texte 160 dans la quatrième. On a choisi de ne pas lui affecter de texte, du fait de la contradiction, mais si on avait choisi le texte par un vote à la majorité, on aurait eu un appariement juste. Sur les 300 appariements possibles, on en a proposé 279 et 278 se sont avérés justes, soit un taux de reconnaissance de 92,67%, avec une seule erreur. Le seul appariement faux est entre le résumé r122 et le texte t51 au lieu du texte attendu t107. Il n'a été proposé que dans une des variantes, comme l'indique la valeur 4 de sa qualité. Le fait qu'il n'y ait aucune erreur dans les associations de qualité 1, 2 et 3 tend à augmenter la confiance qu'on peut avoir dans l'indice de qualité de la comparaison.

Nous venons de voir que la méthode des voisins réciproques a été modifiée pour donner une méthode de prédiction honorable des couples (résumé, texte) attendus, et qu'elle peut être améliorée par combinaison. Mais nous avons laissé de côté des doublons qui peuvent être en partie récupérés. C'est la deuxième étape de notre algorithme d'appariement.

4.2 La deuxième étape de notre méthode d'appariement

Dans l'étape 1 de l'algorithme, nous avons laissé de côté le résumé r90 car il était le plus proche de 295 textes, mais il y a certainement parmi ceux-ci le bon texte à lui apparier, appelons-le t,. Quand nous avons cherché les résumés les plus proches du texte t, le premier était le résumé r90, à la distance d1, et le second, appelons-le r, était à la distance d2 > d1, par construction[1]. Nous supposons que la différence d2-d1 est plus grande pour le texte t que pour tous les 294 textes associés à r90, ce saut important entre les deux résumés indiquant que c'est le résumé r90 qui lui est associé, et non le résumé r. Selon ce principe, nous reprenons tous les doublons et nous créons deux nouvelles listes de couples. Dans notre exemple, la première liste contiendra les couples associés aux résumés r90 et r182, et l'autre les 22 couples correspondant aux 22 textes à l'origine de doublons. Et ces deux listes sont fusionnées de la manière habituelle. Dans notre exemple, un seul des deux couples de la première liste appartient à la seconde, ce qui donne un couple de qualité 3, l'autre disparaît car il est en contradiction avec un de ceux de la liste, et il en reste 20 dans l'autre liste qui sont de qualité 4.

S'il reste encore des textes et résumés qui n'ont pas été appariés, la troisième étape permet d'essayer de les apparier.

4.3 La troisième étape de notre méthode d'appariement

On prend tous les résumés qui n'ont pas été appariés, et on cherche pour chacun le texte correspondant au saut maximum de distance, mais en se limitant aux k plus petites distances (les meilleurs résultats ont été obtenus avec k=10). On créée ainsi une première liste de couples. Puis on procède de la même façon avec chaque texte non apparié pour créer la deuxième liste de couples, et on fusionne les deux listes de la manière habituelle. On contrôle ensuite la cohérence de cette nouvelle liste avec les précédentes, et on ne garde que les nouveaux couples qui ne contredisent pas d'anciens couples. Dans notre exemple, cela n'a produit que 2 nouveaux textes auxquels on a attribué la qualité 4.

5 Résultats dans le cadre du défi (en univers fermé)

On a fait figurer dans le Tableau 3 les résultats obtenus sur les 300 résumés et textes de l'ensemble d'apprentissage en appliquant notre algorithme d'appariement sur les 4 méthodes et leurs variantes, séparément. Il y a une ligne par variante, d'abord les 4 variantes de la méthode de compression, puis la méthode de comparaison utilisant les formes avec les 3 distances, et enfin la méthode utilisant les lemmes avec divers ensembles de catégories de lemmes (1 : mots composés, 2 : noms en dehors du dictionnaire, 3 : adjectifs, 4 : verbes, 7 : noms du dictionnaire). En colonnes on a le nombre de couples justes et faux appariés à chaque étape de l'algorithme, les couples de qualité 1 et 2 de la première étape (Vq1 pour les couples justes de qualité 1, Fq1 pour les couples faux de qualité 1, etc.) , ceux de qualité 3 de la deuxième étape, de qualité 4 de la troisième et le nombre total de couples appariés. On peut voir que tous les couples appariés de qualité 1 (ce sont les voisins réciproques) sont justes, quelle que soit la méthode considérée. On souhaite combiner les méthodes produisant le moins d'erreurs, ce sont donc les méthodes produisant le plus grand nombre de couples de qualité 1. On peut voir en gras les trois meilleures méthodes selon ce critère, et parmi celles-ci, la méthode NeuroNav d'indexation par noms et mots composés, combinée à la distance d'Hellinger, qui donne exactement 300 associations entre résumés et textes, toutes justes, donc 100% de réussite.

[1] On a supposé, pour réaliser cet algorithme, que les distances entre un texte et les résumés n'étaient jamais égales, pas plus que celles entre un résumé et les textes. Mais le fonctionnement de l'algorithme n'est pas gêné par des éventuelles égalités.

Méthode	Paramètres	Distance	Étape 1				Étapes 2 et 3				Total	
			Vq1	Fq1	Vq2	Fq2	Vq3	Fq3	Vq4	Fq4	Vrai	Faux
Compression	res_txt	Dist2	1	0	251	1	1	0	22	0	275	1
Compression	res_txt	Dist1	1	0	124	0	4	0	30	6	159	6
Compression	txt_res	Dist2	1	0	165	0	0	0	52	4	218	4
Compression	txt_res	Dist1	2	0	109	1	3	0	29	1	143	2
Formes		Hellinger	184	0	72	0	28	0	13	0	297	0
Formes		**TF-IDF**	**265**	**0**	**9**	**0**	**14**	**0**	**8**	**0**	**296**	**0**
Formes		Euclidienne	18	0	125	16	11	0	40	18	194	34
Lemmes	**cat 127**	**Hellinger**	**269**	**0**	**5**	**0**	**17**	**0**	**9**	**0**	**300**	**0**
Lemmes	cat 127	TF-IDF	253	0	13	0	21	0	10	0	297	0
Lemmes	cat 127	Euclidienne	220	0	28	1	37	0	10	0	295	1
Lemmes	cat 27	Hellinger	263	0	11	0	18	0	4	0	296	0
Lemmes	cat 27	TF-IDF	252	0	14	0	19	2	7	1	292	3
Lemmes	cat 27	Euclidienne	216	0	27	0	39	1	12	1	294	2
Lemmes	**cat 12347**	**Hellinger**	**265**	**0**	**10**	**0**	**16**	**0**	**7**	**0**	**298**	**0**
Lemmes	cat 12347	TF-IDF	259	0	11	0	19	0	7	0	296	0
Lemmes	cat 12347	Euclidienne	209	0	33	0	40	0	11	2	293	2
Lemmes	cat 2347	Hellinger	257	0	15	0	19	0	7	0	298	0
Lemmes	cat 2347	TF-IDF	259	0	12	0	18	1	5	1	294	2
Lemmes	cat 2347	Euclidienne	210	0	32	0	39	0	12	3	293	3

Tableau 3: Résultats sur l'ensemble d'apprentissage

Dans le Tableau 4, on a fait figurer les mêmes résultats pour l'ensemble Test. On ne connaissait pas le résultat attendu, mais les trois méthodes choisies au moment de l'apprentissage ont donné respectivement 198, 197 et 196 couples sur les 198 attendus. La combinaison des trois méthodes a consisté à contrôler que les 196 couples communs étaient les mêmes, que le 197ème d'une des deux méthodes coïncidait bien avec l'un des deux restants de l'autre méthode, et donc il ne restait plus de doute pour le dernier.

Méthode	Paramètres	Distance	Étape 1				Étapes 2 et 3				Total	
			Vq1	Fq1	Vq2	Fq2	Vq3	Fq3	Vq4	Fq4	Vrai	Faux
Compression	res_txt	Dist2	0	0	29	2	0	0	128	6	157	8
Compression	res_txt	Dist1	0	0	85	0	0	1	24	4	109	5
Compression	txt_res	Dist2	0	0	49	1	0	0	84	0	133	1
Compression	txt_res	Dist1	0	0	82	2	0	1	22	3	104	6
Formes		Hellinger	87	0	83	2	15	0	6	1	191	3
Formes		**TF-IDF**	**171**	**0**	**9**	**0**	**10**	**0**	**8**	**0**	**198**	**0**
Formes		Euclidienne	1	0	32	23	4	0	20	22	57	45
Lemmes	**cat 127**	**Hellinger**	**180**	**0**	**3**	**0**	**9**	**0**	**4**	**0**	**196**	**0**
Lemmes	cat 127	TF-IDF	170	0	8	0	15	0	5	0	198	0
Lemmes	cat 127	Euclidienne	161	0	12	1	14	0	9	1	196	2
Lemmes	cat 27	Hellinger	182	0	2	0	9	0	5	0	198	0
Lemmes	cat 27	TF-IDF	165	0	11	0	17	0	5	0	198	0
Lemmes	cat 27	Euclidienne	161	0	13	1	15	0	5	1	194	2
Lemmes	**cat 12347**	**Hellinger**	**178**	**0**	**6**	**0**	**9**	**0**	**4**	**0**	**197**	**0**
Lemmes	cat 12347	TF-IDF	177	0	5	0	10	0	6	0	198	0
Lemmes	cat 12347	Euclidienne	153	0	21	0	19	0	4	0	197	0
Lemmes	cat 2347	Hellinger	178	0	8	0	8	0	3	0	197	0
Lemmes	cat 2347	TF-IDF	178	0	4	0	9	0	7	0	198	0
Lemmes	cat 2347	Euclidienne	152	0	22	0	19	0	4	0	197	0

Tableau 4 : Résultats sur l'ensemble de test

6 Conclusions, perspectives

Ayant obtenu 0 erreurs pour la piste 2, la plus difficile, il n'était pas étonnant que l'on obtienne le même niveau de performance pour la piste 1 (appariement résumés-textes complets), ce qui nous a amené à la première place ex-aequo du classement. En dehors de cette satisfaction, plusieurs conclusions peuvent être tirées :

- Pour la tâche d'appariement avec des textes amputés de l'introduction et de la conclusion, il n'était pas évident au départ que l'ordinateur pourrait faire au moins aussi bien, et bien plus rapidement, qu'un lecteur intelligent et spécialiste d'un domaine difficile des sciences humaines. C'est pourtant ce qui s'est produit.

- Nous avons mis au point une méthode d'appariement robuste qui serait utilisable sans grandes modifications en univers ouvert (textes plus nombreux que les résumés, certains résumés sans textes, …) en se contentant de retirer les phases de calculs de distances partant des textes vers les résumés. Vers un prochain défi ?

- Les choix qui ont sous-tendu la conception du logiciel NeuroNav sont confirmés : pour les calculs de distance entre textes, l'utilisation des seuls lemmes de substantifs et termes composés ; pour le type de distance, celle de Hellinger plutôt que celle du TF-IDF, aux performances un peu inférieures et peu compatible avec un univers ouvert, ou que la distance euclidienne sur la sphère, bien moins performante. Une augmentation de la qualité de l'analyse morpho-syntaxique, quelque peu rudimentaire aujourd'hui, laisserait espérer des performances voisines de 100% en univers ouvert

- De façon surprenante, la distance de compression donne 93% de réussite sur l'ensemble d'apprentissage, 88,4% sur l'ensemble de test, avec 6 erreurs dont une due au texte vide, performances qui sont loin d'être négligeables pour une méthode aveugle, multilingue et multi-codages, sans aucune intelligence ajoutée ! Des méthodes de compression plus efficaces que celle, ancienne, de déflation pourraient encore augmenter ces scores et rivaliser avec les méthodes morpho-syntaxiques « intelligentes ».

Références

BANERJEE A., DHILLON. I., GHOSH J.AND SRA S. (2005). Clustering on the Unit Hypersphere using Von Mises-Fisher Distributions. *Journal of Machine Learning Research* (JMLR), vol. 6, 1345-1382.

CILIBRASI R. (2003). Clustering by compression. IEEE *Trans. on Information Theory*, 51(4) : 1523–1545.

DELPRAT B., HALLAB M., CADOT M., LELU A. (2011). Processing a Mayan Corpus for Enhancing our Knowledge of Ancient Scripts, *4th. International Conference on Information Systems and Economic Intelligence* (SIIE 2011) February 17, 18 and 19th., 2011, Marrakech (Maroc).

DOMENGES D. ET VOLLE M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, N° 35, P. 3-83.

ESCOFIER B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionnelle. *Revue de Stat. Appliquée*, 26(4):29-37, Paris.

LEGENDRE P., GALLAGHER E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*: 129: 271-280.

LELU A. (2003). Evaluation de trois mesures de similarité utilisées en sciences de l'information. *Information Sciences for Decision Making* 6 14–25.

LELU A., CADOT M., AUBIN S. (2006). Coopération multiniveau d'approches non supervisées et supervisées pour la détection des ruptures thématiques dans les discours présidentiels français. *Semaine du Document Numérique 2006 / DEFT'06*, 21-22/9/2006, Fribourg, Suisse.
http://www.lri.fr/~aze/fdt/DEFT06/articles/DEFT06-Lelu-Cadot-Aubin.pdf

RENYI A. (1966). *Calcul des probabilités*, Paris, Dunod, 620 p.